Self-poised Ensemble Learning*

Ricardo Ñanculef[1], Carlos Valle[1], Héctor Allende[1], and Claudio Moraga[2]

[1] Universidad Federico Santa María,
Departamento de Informática, CP 110-V Valparaíso, Chile
{jnancu, cvalle, hallende}@inf.utfsm.cl
[2] Dortmund Universitaet, 44221 Dortmund, Deutschland
claudio.moraga@udo.edu

Abstract. This paper proposes a new approach to train ensembles of learning machines in a regression context. At each iteration a new learner is added to compensate the error made by the previous learner in the prediction of its training patterns. The algorithm operates directly over values to be predicted by the next machine to retain the ensemble in the target hypothesis and to ensure diversity. We expose a theoretical explanation which clarifies what the method is doing algorithmically and allows to show its stochastic convergence. Finally, experimental results are presented to compare the performance of this algorithm with boosting and bagging in two well-known data sets.

1 Introduction

As long as problems modern data analysis has to tackle on become harder, machine learning tools reveal more important abilities for the process of extracting useful information from data [7].

Intuitively, learning systems are such that they can modify their actual behavior using information about their past behavior and their performance in the environment to achieve a given goal [9]. Mathematically speaking, the supervised learning problem can be put in the following terms: given a sample of the form $T = \{(x_1, y_1) \ldots (x_n, y_n)\}$, $x_i \in X$, $y_i \in Y$ obtained sampling independently a probability measure P over $X \times Y$, we are asked to recover a function $f_0(x)$, such that it minimizes the risk functional

$$R(f) = E_{X \times Y}\left[Q(f(x), y)\right] = \int_{X \times Y} Q(f(x), y) dP(x, y) \qquad (1)$$

where Q is a problem specific loss function, integrable in the measure P. The space X is usually named "the input space" and Y "the output space" of the problem, when the function f_0 -to be recovered- is thought as a mapping from X to Y, underlying the particular sample pairs (x_i, y_i). Since in realistic scenarios, the measure P is not known, the risk functional cannot be computed and

* This work was supported in part by Research Grant Fondecyt (Chile) 1040365 and 7040051, and in part by Research Grant DGIP-UTFSM (Chile). Partial support was also received from Research Grant BMBF (Germany) CHL 03-Z13.

A.F. Famili et al. (Eds.): IDA 2005, LNCS 3646, pp. 272–282, 2005.

neither its minimum f_0. So, it is necessary to develop some induction criterion to be minimized in order to get an approximation to this, which is typically the empirical risk

$$R(f) = \hat{E}_{X\times Y}\left[Q(f(x), y)\right] = \sum_i Q(f(x_i), y_i) \tag{2}$$

Machine learning deals with the problem of proposing and analyzing such inductive criteria. One of the most successful approaches introduced recently in this field, relies on the idea of using a set of simple learners to solve a problem instead of using a complex single one. The term used to describe this set of learning machines is “an ensemble” or “a committee machine”.

The practice and some theoretical results [5] [6] have revealed that diversity is a desirable property in the set of learners to get real advantages of combining predictors. Ensemble learning algorithms can then be analyzed describing the measure of diversity they are - implicitly or explicitly- using and the way in which they are looking for the maximization of this quantity. Boosting and bagging, for example, introduce diversity perturbing the distribution of examples each machine uses to learn. In [14] an algorithm to introduce diversity in neural networks ensembles for regression was described. In this approach we encourage decorrelation between networks adding a penalty term to the loss function. [17] and [5] are examples where negative correlation between learners of the ensemble is looked for.

In the present paper, an algorithm to generate ensembles for regression is examined. It is shown that this algorithm introduces a strictly non-positive correlation between the bias of learner at time t and the average bias of previous learners. This approach allows to introduce diversity without damaging the performance of the machine on the training set. Theory shows that this approach stochastically converges, while experimental results show that this works well in practice. The structure of this paper is as follows: in the next section we present the ensemble model for learning from examples discussing the notion of diversity in a regression context. We also introduce for comparing purposes, boosting and bagging. Following with this idea, we present in the third section the proposed algorithm and discuss the theoretical aspects of the same. Finally we show a set of experiments on two difficult benchmarks, to test the final algorithm and compare the results with boosting and bagging. The fifth section is devoted to some concluding remarks and future work.

2 Ensemble Learning and Diversity

To solve the problem of learning from examples -formalized in the latter section- one needs to choose an hypothesis space H for searching the desired f_0 or an approximation to this. Statistical learning theory establishes that the particular structure of this space is fundamental for guaranteeing the well-behavior of the inductive criterion selected for the learning machine.